Low Voltage SRAMs with Adequate Stability in Nanoscaled CMOS

Jiajing Wang
Department of Electrical and Computer Engineering
University of Virginia

A Dissertation Proposal Presented in Partial Fulfillment of the
Requirement for the
Doctor of Philosophy Degree
in Electrical Engineering

October 26, 2008

Abstract

Increased leakage current and device variability are the major challenges with CMOS technology scaling. Since Static Random Accessed Memory (SRAM) is often the largest component in the embedded digital systems or System-on-Chip (SoC), it is more vulnerable to those challenges. To effectively reduce SRAM leakage and/or active power, supply voltage (V_{DD}) is often scaled down during standby and/or active operation. For ultra-low-energy applications, SRAM is even required to operate with V_{DD} near/below the threshold voltage. However, SRAM stability are weakened under lower voltages. Furthermore, variation, especially mismatch, enormously degrades SRAM stability and yield. Our overall goal is to help designing low voltage SRAM for power/energy reduction while maintaining enough stability and yield in the presence of variations.

We first investigate SRAM standby operation with V_{DD} scaling, and propose a statistical method to fast and accurately predict the worst data retention voltage (DRV), i.e. the minimum V_{DD} that an SRAM can preserve all the data during standby in the presence of variations. Then we propose an adaptive approach to achieve aggressive standby V_{DD} scaling under process, voltage and temperature (PVT) variations. Our approach uses canary replica bitcells with online failure detectors and the feedback controller to an external DC-DC converter so that V_{DD} can be adjusted with PVT changes tracked by canary replicas. Data reliability in aggressive V_{DD} scaling is ensured by a critical failure threshold, which can be programmed for trading off leakage power savings with yield. Silicon results from both 90nm and 45nm test chips will be evaluated. We also propose several techniques to enhance the adaptiveness and automation of the canary system, and thoroughly analyze its effectiveness with considering of overhead sources as well as the technology scaling effect. After that, we will further investigate SRAM yield under lower voltages. With technology scaling, SRAM yield loss due to degraded stability limits the lowering of the minimum V_{DD} in dynamic voltage-and-frequency scaling (DVFS) systems or ultra-low-energy systems. Accurate yield analysis becomes indispensable. Therefore, we propose to estimate yield with accurate and fast analysis of the failure probability under variations for lower supply voltages.

1 Introduction

1.1 Motivation of Low Voltage SRAM

For more than three decades, Moore's law has successfully predicted the scaling trend of process technologies, i.e., the number of transistors per chip doubles roughly every 24 months or the feature of CMOS device scales roughly by 0.7 per generation. This scaling trend has been contributing to the significant improvement of density and performance for IC designs. However, the continuity of technology scaling in sub-100nm region encounters several major challenges. One of them is the increased leakage current.

Leakage current is the current when the transistor is off. It composes of the subthreshold leakage current, the gate leakage current, the pn junction reverse-bias current, and the gate-induced drain leakage current. The subthreshold leakage current is the dominant component, which occurs due to the weak inversion conduction when the gate-source voltage (V_{GS}) is less than the threshold voltage (V_T), and changes exponentially with ($V_{GS} - V_T$). The reduction of V_T with technology scaling causes a substantial increase in the subthreshold leakage current. In recent years, the gate leakage current grows rapidly because the reduced gate oxide thickness results in more electrons tunneling through gate oxide, and it changes exponentially with the gate-drain voltage (V_{GD}) or the gate-source voltage (V_{GS}).

The dramatically increased leakage power becomes a major portion of the total power consumption in scaled technologies, especially for battery operated systems. Because SRAM/Cache often occupies the largest percentage (up to 90%) of the chip area, its leakage power dominates the total leakage power of the chip. Therefore, it is highly demanded to reduce leakage power in SRAM.

Several techniques have been proposed to reduce SRAM leakage current. Dual-V_T [1] and body biasing [2] utilized V_T control to reduce subthreshold leakage current. Source biasing [3, 4] and voltage scaling [5] approaches collapse the actual rail-to-rail voltage of SRAM cells so that both the subthreshold and gate leakage current can be reduced. Because of the advantage of gate leakage reduction, voltage scaling and source biasing would be more effective for leakage power reduction in deeply scaled technologies. In the meantime, dynamic voltage scaling becomes popular for saving dynamic power consumption at lower performance requirement. In recent years, energy efficient and ultra-low-power applications such as wireless sensor networks and implantable biological systems have seen the demands of subthreshold SRAM, which operates with V_{DD} even near/below V_T [6–10]. Therefore, various application needs and CMOS technology scaling challenges demand SRAMs to operate at a lower or ultra-lower voltage in standby and/or active mode.

1.2 Difficulties of Low Voltage SRAM

1.2.1 Stability & Yield

Besides performance and power, stability and yield are the key ingredients of SRAM design. There are four failure mechanisms associated with SRAM stability and yield due to parametric variability: hold failure, read failure, write failure and access failure. Hold failure occurs if the cell doesn't have adequate noise margin to preserve data in the occurrence of noise sources like radiations and alpha particles. Read failure is caused by the disturbance on the node holding '0' upon read and results in the flipping of data after read. Write failure is mainly caused by the incapability of pulling down the node initially holding '1' and thus the cell data is unable to toggle during write. These three categories of failures are mainly determined by SRAM cell stability. The access failure occurs when an insufficient voltage difference is developed for read sensing within the required timing period, so it is also impacted by the performance of the sensing scheme (e.g. offset of sense amplifier).

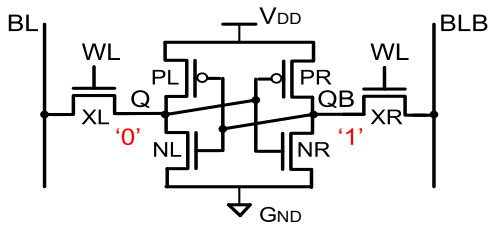

Figure 1: A standard 6T bitcell.

Typically SRAM is designed to have enough hold, read and write noise margins as well as access speed for the required yield under the nominal supply voltage. However, supply voltage would be lowered for the sake of power savings during either standby or active operation. The drawback of supply voltage scaling is the degradation of on current and all kinds of noise margins. It is important to first understand the degradation of stability and yield with supply voltage scaling. Then we can find the minimum supply voltage (Vmin) SRAM array can operate with the required yield and thus achieve the best power savings. Also it could help designers to improve stability and yield under lowered supply voltages so that the Vmin limit can be lowered for more power savings.

1.2.2 Variability

Variation has become the biggest challenge with technology scaling. From the aspect of the impact region, all the variations can be categorized into two groups, the global ones and the local/random ones. Global variations include the manufacture related process variation, voltage supply fluctuation, and temperature change (i.e. PVT variations). Random Doping Fluctuation (RDF) and line edge roughness are the major sources of the random variation for V_T. And RDF induced variation becomes worse with scaling. The randomness of V_T due to RDF can be modeled as a normal distribution with the standard deviation inversely proportional to the channel area [11]. SRAM commonly uses symmetrical transistors with the smallest geometry for high density, thus it naturally suffers more random variation or mismatch than the logic designs. Moreover, for big SRAMs with Mega-bit capacity, $\geq 5\sigma$ points of variation have to be evaluated in order to meet the yield requirement.

Variation impacts all the SRAM metrics. On current, leakage current and noise margin are all changed with variations. Since leakage current changes exponentially with V_T, small V_T variation could cause a huge increase of leakage current. Because of using the minimum-geometry cell , random variation/mismatch has a huge impact on SRAM noise margin and yield [12, 13]. PVT toleration and compensation techniques such as body biasing and adaptive sleep transistor design [14, 15] have been proposed to improve power savings and yield. Furthermore, variation impact is exaggerated under lower voltages. In recent years, different techniques have been proposed to tolerate V_T fluctuations and improve read stability and write ability for lowering SRAM operation voltage, such as voltage control on cell supply/source voltage, wordline voltage and bitline voltage as well as using additional transistors [16–20].

1.3 Goals

Our work will address the major difficulties of low-voltage SRAM design in deep sub-100nm technologies. We will mainly focus on the standard 6T SRAM bitcell as shown in Fig.1. The major goals of our work are:

Problem statement: With technology scaling, low voltage SRAM operation is highly demanded for leakage/dynamic power reduction. The minimum operation voltage of SRAM is mainly limited by the yield loss due to stability degradation in the presence of variation.

- Investigate hold stability under lower voltage, and propose a statistical method to fast and accurately predict the minimum supply voltage that SRAM can operate with during standby in the presence of variation.
- Propose an adaptive scheme to achieve aggressive standby V_{DD} scaling under PVT variations while maintaining the required yield.
- Analyze the failure probability of SRAM under lower voltage and estimate the minimum operational voltage for the required yield.

2 Statistical Analysis for Data Retention Voltage

2.1 Motivation

2.1.1 Standby V_{DD} Scaling and Data Retention Voltage

Supply voltage scaling has been used to reduce SRAM standby leakage power since the reduced V_{DD} decreases subthreshold leakage, gate leakage as well as junction leakage current [21, 22]. However, the collapsed signal rail ($V_{DD} - V_{GND}$) degrades the cell's stability. Fig. 2 shows excessive V_{DD} scaling causes the cell losing its original data ('0' in this example). Fig. 2(a) shows the case when the cell is balanced (without mismatch). The cell nodes Q and QB converge to a metastable point as a result of degraded gain. Fig. 2(b) shows the case when the cell is imbalanced by some mismatch. In this case, Q and QB flip to the more stable state ('1' here). Data retention voltage (DRV) defines the minimum V_{DD} that can be applied on an SRAM cell without losing data. It should be noted that the imbalanced cell has a much higher DRV than the balanced one. If we name the DRV for holding '0' as DRV0 and the DRV for holding '1' as DRV1, then the actual DRV should be max(DRV0,DRV1).

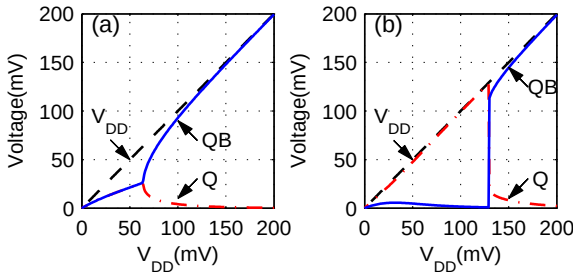

Figure 2: Cell nodes Q and QB either converge or flip when V_{DD} is lowered than DRV.

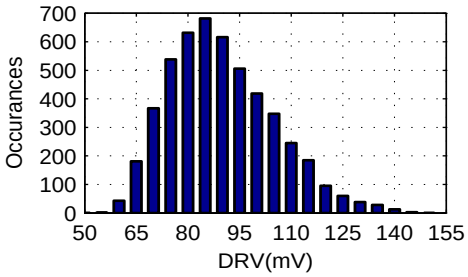

Figure 3: Simulated DRV histogram for a 5K-b SRAM in a 90nm node.

2.1.2 Variation Impact on DRV

Within-die device variation (i.e. mismatch) is the major source causing the cells imbalanced. Therefore, DRV of the individual cell in a memory array exhibits a statistical characteristics. Monte-Carlo (M-C) simulation is a well-known approach that can estimate DRV distribution for an SRAM array given the array size and the statistical parameters of the device variation. The normally-distributed V_T variation caused by random dopant fluctuation is the most dominant component of parametric variation. So we mainly consider its impact on DRV. Fig. 3 is the histogram of a 5k-point M-C simulation showing the DRV for SRAM bitcells in a commercial 90nm CMOS node. It is obvious that DRV is a non-Gaussian distribution with a longer tail at the right side. The tail point is the most critical value since it determines the minimum standby V_{DD} that can be applied on the fault-free SRAM.

2.1.3 Related Work

The most straightforward method for obtaining DRV tail value is to run full M-C simulation. However, a large scale of M-C simulation for memories with multi-million bits can take huge amounts of time. The second way is to run M-C simulations on small arrays and then use extrapolation to model the tail for large-memories. But since DRV is a non-Gaussian distribution, the simply extrapolation with mean and sigma to estimate large sigma point values is inaccurate.

For statistical M-C simulations like DRV, only the samples on the tail are highly interested, so it is unnecessary to run all of the samples except those from far out the tail. Statistical Blockade (SB) tool [23] was proposed to improve upon traditional M-C for simulating rare events. To reduce simulation time, the Blockade tool classifies the possible M-C samples prior to simulation and selects only a subset of them that are likely to appear on the tail for simulation. After simulating this subset of points, the tool identifies the true tail points and uses them to fit a Generalized Pareto Distribution (GPD) model thus allowing estimation of events even farther out in the tail.

Besides fast monte-Carlo simulation approach, accurate and fast modeling is also appealing. [24] first proposed a theoretical model for a cell's DRV. While with occurrence of random variations, a statistical model for DRV estimation especially at tail is more desired. Therefore, we will propose a fast and accurate statistical model to predict the tail of DRV.

2.2 Proposed Research

2.2.1 Statistical DRV Model

Since DRV is the minimum V_{DD} that a cell can preserve its data, we can consider it as the V_{DD} at which static noise margin (SNM) is equal to zero in a noiseless system. Therefore, we propose to use SNM as a starting point to explore DRV statistics.

The most well-known metric for SRAM noise margin is the butterfly curve based SNM, which measures the maximum amount of voltage noise that a cell can tolerate [25]. Particularly, we call SNM for holding '1' as SNMH and for holding '0' as SNML. The actual SNM is the minimum of SNMH and SNML. M-C simulation results have shown that SNMH and SNML are approximately identical normal distributions under the normal distributed V_T variation caused by RDF. If assuming their independence, according to order statistics, the real SNM CDF/PDF can be calculated with the minimum model, which actually shows a good estimation at the tail points [26]. However, questions like how the SNM statistics changes with V_{DD} scaling and whether the minimum model is still valid for lower V_{DD} are not addressed. For DRV statistics,

those answers become critical. Interestingly, we found out that the mean of SNMH/SNML moves with V_{DD} while their standard deviation keeps same. In addition, the mean is approximately linear with V_{DD}, and this linearity doesn't change with variations. So if we have the mean (μ_0) and the standard deviation (σ_0) of SNMH/SNML at one V_{DD} point (V_0) and the slope (k) of the mean versus V_{DD}, we can derive the SNM CDF at any V_{DD} point. Since DRV is the V_{DD} when SNM=0, we can derive the CDF model and inverse CDF model of DRV as shown in (1) and (2).

$$F_{DRV}(x) = 1 - erfc\left(\frac{\mu_0 + k(x - V_0)}{\sqrt{2}\sigma_0}\right) + \frac{1}{4}\left(erfc\left(\frac{\mu_0 + k(x - V_0)}{\sqrt{2}\sigma_0}\right)\right)^2 \quad (1)$$

$$F_{DRV}{}^{-1}(x) = \frac{1}{k}\left(\sqrt{2}\sigma_0 \cdot erfc^{-1}\left(2 - 2\sqrt{x}\right) - \mu_0\right) + V_0 \quad (2)$$

where $erfc(\cdot)$ and $erfc^{-1}(\cdot)$ are the complementary error function and its inverse function. k, μ_0 and σ_0 are fitting coefficients; k can be extracted from a DC sweep simulation and μ_0 and σ_0 can be extracted from a small-scale (1.5K-5K) Monte-Carlo simulation [27].

2.2.2 Model Evaluation

We use an industrial 90nm technology to evaluate our new DRV model. Since we are only interested in the tail of DRV, we propose to compare our model with Blockade tool as well as M-C results in term of the worst DRV value for a fault-free memory with a given size. Fig. 4 shows the comparison result. The size of the memory is represented by the corresponding sigma value (e.g. $\sim 6.1\sigma$ stands for a 1G-b memory). Results from (2) closely track the M-C results with an average error of 1.3% out to 6σ. M-C points greater than 5σ were simulated using the selected points from the Blockade tool classifier, thus allowing dramatically reduced simulation time. The GPD model produced by the Blockade tool also closely matches the M-C data with an average error of 1.0%. In addition, the two models match each other at the tail even far out to $7 \sim 8\sigma$, which is too time-consuming for using filtered Monte-Carlo. We also show the estimation from Normal and Log-normal models that were based on a 5k-point M-C simulation for comparison. The Normal model underestimates DRV while the Log-normal model overestimates it. We should note that our model is not accurate for smaller sigma value points or mean value of DRV (when $\sigma \leq 0$) because SNMH and SNML are actually negatively correlated. However, the intention of our model is to predict DRV tail value for large SRAMs, which is more important for determining the lowest V_{DD} that the SRAM can operate during standby.

3 Canary System for SRAM Aggressive Standby Power Reduction

3.1 Motivation

Fig. 5 shows the DRV distribution of a 5-Kb SRAM array at three global cases (typical/best-case/worst-case). Each case shows a 5K-point M-C simulation with within-die V_T variation plus certain global variations, including process variation, temperature change and voltage fluctuation (PVT variations). It is obvious that for each global scenario, local variations spread the DRV of the cells across the same array, and the cell with the largest DRV (the tail marked with circle) actually determines the minimum standby V_{DD} (Vmin) that can be applied to the whole SRAM array under the current global variations. In contrast, the global variations predominantly move the entire DRV distribution (and the tail) around, so the Vmin for the whole SRAM shifts with global effects.

To account for this variability, existing V_{DD} scaling approaches add a safety margin to the worst scenario

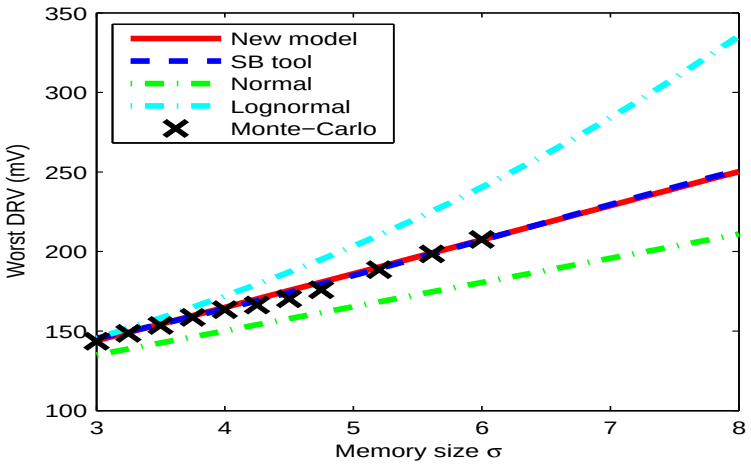

Figure 4: DRV Model Comparison

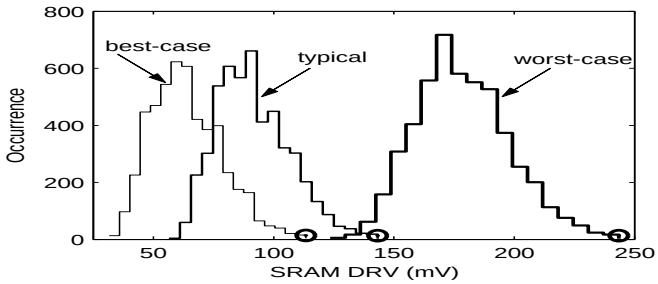

Figure 5: Simulated DRV distribution of a 5-Kb SRAM under typical, best-case and worse-case PVT variations.

to prevent the loss of data. Many previous works select a standby V_{DD} at design time that maintains sufficient margin to protect data in the cells (e.g., the drowsy cache in [28] and the microprocessor with a drowsy mode in [29]). This open-loop approach can leave substantial power savings on the table because the full range of potential DRVs can be quite large when accounting for the worst case. With the scaling of technology, we can expect to sacrifice more leakage power savings by using this conservative worst-case approach due to increased device variability.

Closed-loop control of standby V_{DD} offers an appealing alternative for conditions that allow extra power savings without data loss. Therefore, we propose a feedback architecture using canary replicas for SRAM bitcells [30].

3.2 Canary Adaptive System

Fig.6(a) shows the proposed feedback loop used to lower V_{DD} for leakage power savings while protecting data by keeping V_{DD} above the DRV for the core cells. An on-chip or off-chip voltage regulator supplies V_{DD} to the core cells and to the canary replicas. Multiple canary categories are designed to fail across a range of voltages above the DRV of the SRAM cells as illustrated in Fig.6(b) and maintain this behavior despite changes in global variations and environmental conditions. Local variation smears the distribution of canary DRVs in each set, but the canary distributions are not good indicators of the core cell distribution because there are too few canary cells. We will emphasize that the purpose of the canary categories is not to estimate the full distribution of the core SRAM cells, but instead to sense the proximity of the currently

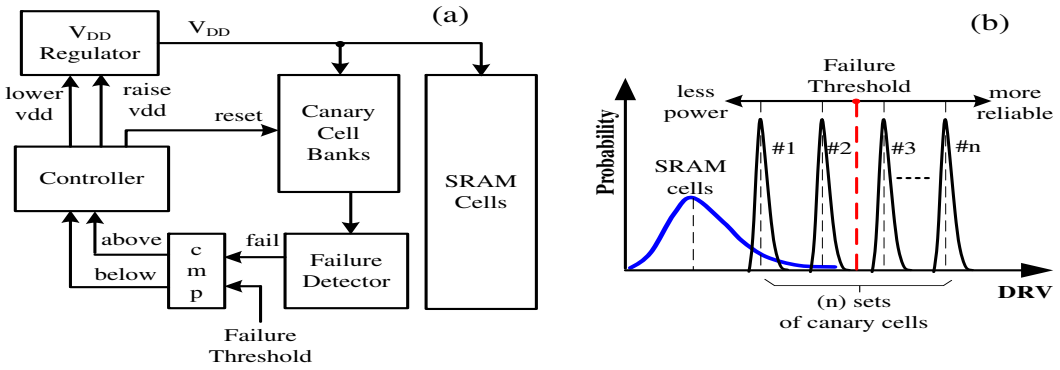

Figure 6: Canary system (a) close-loop structure and (b) mechanism.

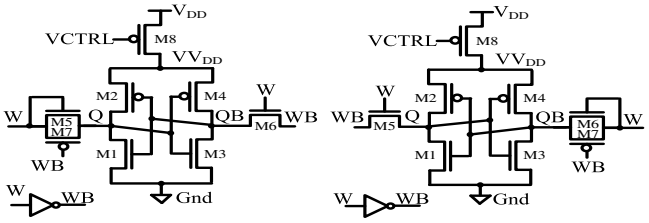

Figure 7: Circuit for (a) Canary cell '0' and (b) Canary cell '1'.

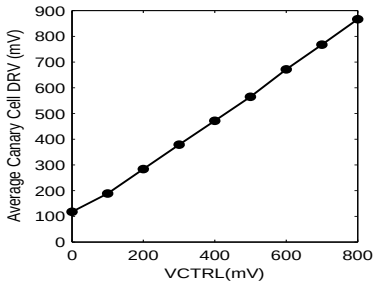

Figure 8: Measured canary DRV vs. VCTRL.

applied VDD to the worst DRV of the SRAM cells. The failures of canary categories are monitored by the online failure detectors. SRAM data safety is ensured by a programmable failure threshold, which defines the critical failure status of each canary category and determines the proximity of the applied standby V_{DD} to the tail of the SRAM DRV distribution. When entering the standby mode, the controller starts lowering V_{DD} until the canary failures meet the failure threshold. Once the global stimuli occur, the canary failures will exceed or drop below the failure threshold, which triggers the controller to raise or lower V_{DD} accordingly.

Another advantage of this approach is to allow a tradeoff between power savings and data reliability by altering the failure threshold. When the application needs a higher data reliability, a higher failure threshold should be chosen. Instead, when data reliability requirement is lowered or some date errors can be tolerated by redundancy or error correction techniques, the failure threshold can be lowered for more power reduction.

3.3 Proposed Research

The most important component in the canary system is the canary cell. We proposed the circuit in Fig.7 (a) and (b) as canary cells for holding '1' and '0', respectively. Each canary cell contains the same 6T transistors (M1∼M6) as any SRAM cell. To enhance the write capability at sub-threshold supply voltages (e.g. for canary reset), another PMOS pass transistor (M7) is added to the side of the cell that stores a '1'. The input signal, W, and its inversion, WB, act as the bitlines and wordline for writing data to the cells during a reset. When W is high, the canary cell resets its data; when it is low, the canary cell enters the standby mode. A PMOS header (M8) is inserted between the supply voltage of the canary cell and VDD, and another input signal VCTRL drives its gate. These small circuit modifications, especially the add of PMOS header, contribute to a wide-range tuning of DRV for canary cell.

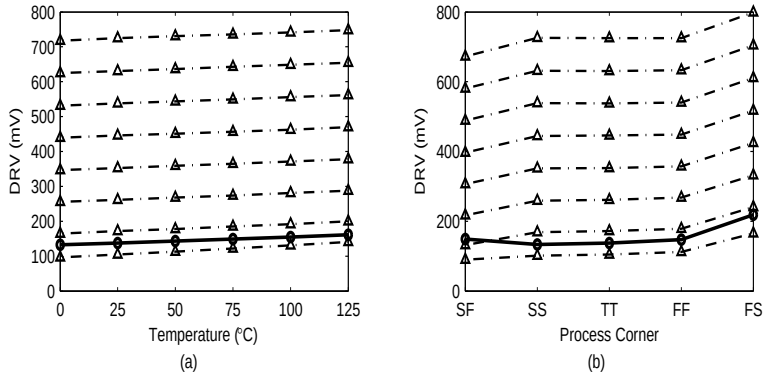

Figure 9: Canary tracks (a) temperature change and (b) process variation.

A 90nm CMOS bulk test chip implements all of the circuits in Fig. 6(a) except the V_{DD} regulator. Fig. 8 shows the measured average DRV of canary cells versus VCTRL at room temperature. The control of the header allows us to provide the desired continuum of failure voltages for the canary cell. It also verifies an approximately linear relationship between the canary DRV and VCTRL, so canary DRVs can be placed at regular intervals above the core DRV with evenly spaced VCTRLs.

In addition, the canary's capability of tracking PVT variation is investigated. Fig. 9(a) and (b) show simulated results that compare the canary behavior with an SRAM array across temperature changes and global process corners. We used a 1-Kb SRAM as an example. The curve with circles stands for the worst DRV of the 1-Kb SRAM, and the curves with triangles stand for different canary sets (the upper ones are the sets with higher VCTRL). The upper 7 canary sets consistently fail before the SRAM at all the temperatures and all the process corners with the only exception of the SF (Slow-N Fast-P) corner. This indicates that the canaries will successfully track global effects on the SRAM array.

Based on the results of the prototype, we propose several techniques to enhance the adaptiveness and automation of the canary system. We will also propose to thoroughly analyze the effectiveness of our scheme with considering of overhead sources as well as the technology scaling effect.

3.3.1 Canary Cell Improvement

In the first prototype, we proposed that each canary cell only contains one stand-alone 6T bitcell, the PMOS header for tuning the actual supply voltage of the 6T bitcell, and an additional PMOS pass transistor for enhancing write ability at lower supply voltage. To improve the correlation of global effects on canary cells and core cells, we propose a new structure for the canary cell. We add dummy bitcells around the functional bitcell to mimic the real physical environment of an SRAM cell. To reduce area cost, a mini 3x3 SRAM array will be used. The central bitcell of the mini-array will be monitored by the failure detector. To ensure the canary cell behaves more like SRAM cells in the presence of the global variations, we will use the same layout pattern as the real SRAM array except some minor changes on metal wires for pulling out the storage nodes of the central cell. The power supply of the mini array will be connected with the PMOS header so that the canary cell can fail at a higher VDD than the worst SRAM cell.

Previously, for each canary category, we used two separate canary cells, cancell-0 and cancell-1, which are designed particularly for holding '0' and '1' individually. One canary category fails when either its cancell-0 or cancell-1 fails. Although this way is simple and easy to implement, it has one drawback. Mismatch causes a cell to be much more stable at one data value than the other, and it is uncertain which data value is more stable due to randomness of local variation. For one canary category, if both cancell-0

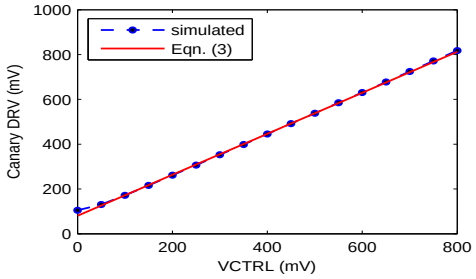

Figure 10: Estimated canary DRV from (3) vs. VCTRL compared with the simulated results.

and cancell-1 happen to be more stable at the value that they are holding, this canary category will never fail or fail at a very low supply voltage regardless of the VCTRL value. We observed this issue in our test chip. To eliminate this possibility, we will propose a new circuit to self-reset the canary cell with its less-stable data value

3.3.2 Model to Estimate VCTRL

As shown in Fig. 8, the canary DRV changes approximately linearly with VCTRL. This linear dependency can be modeled by analyzing the current through the PMOS header M8 (in Fig. 7(a)). We can assume the minimum current for holding the cell data is unchanged with VCTRL. This is reasonable because the minimum current is mainly determined by the variations on the cross-coupled inverters. The minimum current through M8 should also remain constant, and thus an equation to estimate the required V_{DD} for holding cell data vs. VCTRL can be derived as (3).

$$V_{DD} = \frac{VCTRL}{1+\eta_8} + \frac{n_8 V_{th} ln(K)}{1+\eta_8}. \tag{3}$$

where η_8 is the DIBL coefficient of M8 and n_8 is its subthreshold swing factor, V_{th} is the thermal voltage, and K is a constant fitting parameter. Fig.10 compares the canary DRV values from (3) with the simulated results. This first-order linear model provides a good approximation for most VCTRL values. By combining (2) and (3), we can estimate the VCTRL value necessary to satisfy a given SRAM reliability constraint with (4):

$$VCTRL = \frac{1+\eta_8}{k}\left[\sqrt{2}\sigma_0 \cdot erfc^{-1}\left(2-2\sqrt{x}\right) - \mu_0\right] + V_0 \cdot (1+\eta_8) - n_8 V_{th} ln(K). \tag{4}$$

where x is the probability that SRAM DRV is less than the canary DRV with the desired VCTRL value, and all the other parameters are the same as in (2) and (3).

3.3.3 Self Calibration

The power of our canary scheme is the PVT-independent failure threshold, which determines the proximity of the applied V_{DD} to the tail of the SRAM DRV as illustrated in Fig.6(b). The proximity implies the stability margin for V_{DD} scaling. Therefore, the required stability will maintain same under all the PVT variations. It also means that we can simply configure the failure threshold according to the required power and reliability constraints on the normal condition. However, local randomness might cause the tail of SRAM DRV differing from one die to another. So we have to find a suitable failure threshold for each die. In addition, the failure threshold might be different because of the local spread of canary categories. Techniques to reduce canary local variation should be applied. To automate the calibration of the failure threshold, we

thus propose to incorporate a build-in-self-test (BIST) block.

BIST - Self-calibrate the SRAM DRV Tail (Vmin) First, we propose to design a BIST to particularly calibrate the tail of SRAM DRV (Vmin). Typically, Vmin is searched with V_{DD} scaling (from the nominal voltage to lower voltages) similarly as in [31], where the optimum source voltage for reducing standby power with source biasing is searched from the nominal source voltage 0 to higher values. However, the cell DRV is more likely to be below half of the nominal voltage. To save test time, we propose to search V_{DD} in the ascending order from a lower value that is worse for cell stability.

For each search iteration, we propose to use the simplest scan test - $\{W0 \Uparrow, R0 \Uparrow, W1 \Uparrow, R1 \Uparrow\}$, in which standby operation is inserted between write0(1) and read0(1) operation. During read/write operation, the supply voltage will be switched to the nominal high value. The standby supply voltage will be generated from a voltage regulator, whose inputs will be digital signals controlled by BIST. It should be noted that a sufficient long standby period is needed to enable the cell flip to the opposite state under low supply voltage.

Row/column redundancy and ECC are conventionally used for reducing the yield loss due to manufacturing defects and soft errors. For low standby power operation, they can also been used to tolerate data-retention errors so that the minimum standby voltage can be less than the worst DRV in the SRAM [32]. To accommodate to the redundancy techniques for ultra-low-standby-operation applications, we also propose to implement the BIST with considering the use of those redundancy techniques.

Self-calibrate the Initial Failure Threshold Simulation and measured results have shown that the DRV of the canary cell is approximately linear with VCTRL value. Hence, we can use a series of VCTRL values to create a group of canary categories that fail at regular intervals across a wide range as desired in Fig.6(b). We propose to use a resistor ladder with N distinct voltage nodes to generate the series of VCTRL values. The top of the resistor ladder can be connected with an external reference voltage VREF. So the VCTRL value of the *ith* canary is $VCTRL_i = (i/N) \times VREF$. After finding the V_{DD} closest to the actual SRAM DRV tail (Vmin), the BIST will apply this V_{DD} as the supply voltage for the canary cells and measure the failure status of each canary category. This measured failure status can be used to determine the initial failure threshold suitable for the chip, and this information can be stored in an eFuse or NVM on the die.

3.3.4 Efficiency Analysis

Overhead Analysis In the previous discussions, we have analyzed the benefits of using canary-based V_{DD} scaling without accounting for overhead. We propose to quantify the impact of various overhead sources on the potential savings achievable by our scheme.

The major overhead sources are:

- Canary Circuit
  The canary power overhead includes the power of canary array and peripheral circuits (including all the failure detectors, controller, and other components). The dynamic power of the canary circuits is small relative to their leakage power since the canary system works at a very low frequency. We can thus consider the total overhead power to equal the leakage power of the canary circuits. Since different size of SRAM macro can use the same canary circuit, we will analyze the canary overhead with varied SRAM macro size.
- Voltage regulator (DC-DC converter)
  Our canary-based V_{DD} scaling approach requires a DC-DC converter that can provide a standby supply voltage across a fairly large range of values. Since this low, variable voltage is only supplied during

standby, the load current may be relatively low. The DC-DC regulator may be on-chip or off-chip, but either way, we need to account for the impact of its efficiency on the overall power savings. [33] has described a switched DC-DC converter that can deliver load voltages ranging from 0.3V to 1.1V. That particular converter provided >70% efficiency over a wide range above 0.45V. The minimum efficiency remained larger than 55%. But for conservative analysis, we would also evaluate even lower efficiency values to see whether our scheme strongly relies on the performance of DC-DC converter.

- BIST
  The power overhead of BIST will not be a concern since it would only be used during test time and can be completely turned off during normal operation. But test time cost is usually a concern for a BIST, we will analyze the test time used for SRAM DRV tail calibration.

Scaling Effect We propose to use simulations with predictive technology models (PTMs) [34] to assess the effectiveness of our canary scheme for nodes beyond 45nm. Since the current PTM model doesn't support statistical analysis, we will first add the appropriate V_T variation into the PTM models. Then we will use the variation-enabled model to investigate the SRAM DRV distribution (especially tail) with the technology scaling. We will also evaluate whether canary DRV is still linear with VCTRL under smaller technologies. We will evaluate both the SRAM DRV tail and the canary DRV with temperature changes. To evaluate the process variation impact, we will first add appropriate process corner for each PTM model. Then we will evaluate whether canaries can still track process variations under small technologies.

3.3.5 45nm Test chip

To physically verify our canary scheme with more deeply scaled technologies, we have implemented it on the 2nd test chip with a 45nm bulk technology. We incorporated several improvements including using dummy cells for canary cell on this test chip. The silicon results will be measured and compared with 90nm chip results.

4 Yield Analysis for Low-voltage SRAM

4.1 Motivation

In dynamic voltage-and-frequency scaling (DVFS) system, low-voltage active operation is demanded for saving dynamic power consumption. However, lower-voltage read/write operation is more prone to fail. VCCmin defines the minimum voltage SRAM can operate for the required yield and performance constraint. With technology scaling, yield becomes the major obstacle that limits VCCmin because of the increased parametric variation. For ultra-low-energy applications, minimum energy (minE) and the optimum VDD point (VDDopt) for achieving minE have been explored [35, 36]. Theoretical models have shown that VDDopt often occurs in the subthreshold region, i.e. VDDopt$< V_T$. However, with increased variations, functional yield cannot be met with that VDDopt value. Hence, the actual VDDopt as well as minE are determined by the minimum VDD for the required yield.

To enable lowering of VCCmin or VDDopt for more power/energy reduction in DVFS systems or ultra-low energy systems, yield improvement techniques are often needed. To evaluate their efficiency, an accurate yield analysis method is essential. Since yield analysis is involved with the statistical events, traditional M-C method is often very slow for large scale circuits like SRAM. So a fast yield analysis method is desired.

4.2 SRAM Failure Mechanisms

In SRAM, stability degradation and yield loss are mainly caused by the four failure mechanisms.

- Hold Failure (HF)
  As we discussed in section 2, hold failures occur when the internal node Q and QB flip or converge during the standby mode.
- Read Failure (RF)
  During read, the access transistors (XL and XR in Fig.1) are turned on, as a result the voltage of the node Q (initially holding '0') is bumped up. Read instability occurs when the bumped voltage is too high thus flipping the cell data. To enhance read stability, the driving transistor (NL/NR) is usually designed to be stronger than XL/XR.
- Write Failure (WF)
  During write, one bitline is pulled low and the other is pulled high. To make a successful write, the access transistor (XR) with node '1' must be strong enough to overcome the pull-up transistor (PR) so that the node can be pulled down to '0'. To ensure write ability, XL/XR is usually designed to be stronger than PL/PR.
- Access Failure (AF)
  During read access, one bitline is discharged by the on current through the accessed cell, and the other bitline might also be discharged by the leakage current through all of the unaccessed cells, which is called bitline leakage current. Access failure is more likely to happen when bitline leakage current is comparable to the on current, which results in an insufficient differential input relative to the offset value of the sense amplifier (SA) required for correct sensing.

4.3 Proposed Research

Several literatures have proposed the methods of statistical failure analysis for SRAM in the presence of random parametric variations. [12] proposed stochastic models for butterfly-based read SNM. [13] addressed the importance of statistical simulations to assure SRAM yield. Authors in [37] presented complete and complex models that require numerical solutions to estimate the failure probability values. [38] proposed to model read noise margin and the inverse of write and access delays as Gaussian distribution to perform yield analysis. Statistical methods for yield particularly associated with access failure were also proposed [39, 40]. However, there are several limitations in the existing works. We will address them and propose an accurate yield analysis method.

4.3.1 Failure Probability Definition

If F0/F1 is the event that the cell holding a '0'/'1' fails, all the existing works considered the failure probability $P_F = P(F0) = P(F1)$. However, more strictly speaking, one cell would be considered to fail when either of the scenarios fails, so the actual failure probability should be $P_F = P(F0 \cup F1) = 1 - P(\overline{F0} \cap \overline{F1})$. It should be noted that F0 and F1 are correlated events.

4.3.2 Failure Measurement

Accurate yield estimation relies on correct measurement of each failure category. Since more than one measurement method exist for some failure categories, we will examine their accuracy and complexity, and then choose the best one to use.

Hold failure probability can be evaluated with: $P_{HF} = P(DRV > VDD) = 1 - F_{DRV}(VDD)$. We can apply our DRV CDF model (1) to estimate it.

For read and write failures, there are both dynamic and static methods. Dynamic methods analyze the dependence of the read/write failure on timing [37, 41], so they are usually more accurate. However, because of the complexity of dynamic methods, static ways are often used instead such as butterfly SNM [42], Ncurve metrics [43, 44], BL metric [45] and WL metric [46]. We propose to examine the static methods by comparing them with dynamic margin. If one static way can correctly predict all the failures indicated by the dynamic way, then we can safely use it for faster analysis without losing accuracy. For better accuracy, we also propose to define new dynamic metrics based on the concept of separatrix, which is the boundary of attraction region for the cell state settling to the stable point [47]. Thus a write failure can be defined as the required WL pulse width (T_{WL}) is less than T_{CRITW}, the minimum duration of WL pulse that allows the cell state trajectory cross over the separatrix line during write. On the contrary, a read failure then occurs when $T_{WL} > T_{CRITR}$ during read. The state trajectory can be analyzed by considering read/write events as the current noises injected to the cell nodes.

For access failure, we will evaluate ΔV_{BL}, the differential voltage between the two bitlines within the required access time T_{ac}. ΔV_{BL} is dependent on both the read current and the bitline leakage current. We will also analyze the statistics of SA offset value (V_{OS}). The access failure probability can be expressed as $P_{AF} = P(\Delta V_{BL} < V_{OS})$.

4.3.3 Joint Failure Probability

The overall failure probability of the SRAM is denoted as P_F.

$$\begin{aligned} P_F \quad = \quad & P(HF) + P(RF) + P(WF) + P(AF) \\ & - P(HF\&RF) - P(HF\&WF) - P(HF\&AF) - P(RF\&WF) - P(RF\&AF) - P(WF\&AF) \\ & + P(HF\&RF\&WF) + P(HF\&RF\&AF) + P(HF\&WF\&AF) + P(RF\&WF\&AF) \\ & - P(HF\&RF\&WF\&AF) \end{aligned} \tag{5}$$

The correlation of failure events has been mentioned in [37, 38]. Table.1 lists the possible space of each cell transistor for having the particular failure event. 'S' represents the space of $\Delta V_T < 0$, where the transistor becomes stronger; on the contrary, 'W' represents the space of $\Delta V_T > 0$, where the transistor becomes weaker. It's clear that some of the failure events share one or more spaces (especially HF and RF), which implies the possibility of an non-zero value for the joint PDFs in (5). Therefore, the correlation of the failure events should be analyzed for accuracy.

Table 1: Spaces for the possible failures for cell '0'

Event	XL	NL	PL	XR	NR	PR
HF	S	W	S	W	S	W
RF	S	W	S	W	S	W
WF	W	S	W	W	W	S
AF	W	W	-	-	-	-

4.3.4 Sensitivity to V_{DD}

Although [37] briefly mentioned the sensitivity of failure probability to V_{DD}, it hasn't given the exact relationship between the failure probability and V_{DD}, which is actually critical for determining the minimum V_{DD} with the required yield. We can repeat the same method to estimate the yield at lower supply voltages,

but the computation or simulation for searching the minimum V_{DD} under a given yield constraint could be time-consuming. Therefore, we propose to thoroughly evaluate the sensitivity of each failure probability as well as the overall failure probability to V_{DD}, and explore the method for a fast estimation of minimum V_{DD}.

4.3.5 Evaluation with Predictive Technology Models

We will analyze SRAM yield with predictive models down to 22nm [34] to examine the technology scaling impact on SRAM yield as well as the minimum voltage that SRAM can operate.

5 Research Tasks

Tasks and status of each research goal are listed in Table.2.

Table 2: Research Tasks and Status

TASK	#	DESCRIPTION	OUTCOME	STATUS
DRV Model	1	Model derivation	DRV CDF & inverse CDF model	Done
	2	Model evaluation	Accuracy and speed comparison with M-C and SB tool	Done
Canary	1	90nm test chip	Simulation	Done
			Chip implementation	Done
			Chip measurement	Done
	2	Canary cell improvement	New canary structure and reset circuit	Done
	3	Model for VCTRL	Estimate VCTRL value for a required SRAM reliability constraint	Done
	4	BISTaT	BIST for calibrating SRAM DRV tail	Done
			Self-tune failure threshold	Done
	5	Efficiency analysis	Overhead analysis	Done
			Scaling effect	*Finish in Nov*
	6	45nm test chip	Simulation	Done
			Chip implementation	Done
			Chip measurement	*Finish in Dec*
Yield Analysis	1	HF Prob. estimation	Use our DRV CDF model (1)	Done
	2	WF Prob. estimation	Propose new dynamic write margin. Examine static methods with dynamic one	Done
	3	RF Prob. estimation	Propose new dynamic read margin. Examine static methods with dynamic one	*Finish in Dec*
	4	AF Prob. estimation	Statistically analyze on-current, bitline leakage current, ΔV_{BL}, and SA offset	*Finish in Dec*
	5	HF/RF/WF/AF failure correlation	Obtain the joint failure prob.	*Finish in Jan*
	6	VDD scaling effect	Obtain the failure probability for lower VDDs	*Finish in Jan*
	7	Technology scaling effect	Estimate yield for technologies beyond 45nm	*Finish in Jan*
Writing up	1	Thesis writing	Final Dissertation	*Finish in May*

References

[1] Y. Ye, M. Khellah, D. Somasekhar, A. Farhang, and V. De, "A 6-ghz 16-kb l1 cache in a 100-nm dual-vt technology using a bitline leakage reduction (blr) technique," *IEEE J. Solid-State Circuits*, vol. 38, no. 5, pp. 839–842, 2003.

[2] C. Kim, J.-J. Kim, S. Mukhopadhyay, and K. Roy, "A forward body-biased low-leakage sram cache: device, circuit and architecture considerations," *IEEE Trans. VLSI Syst.*, vol. 13, no. 3, pp. 349–357, March 2005.

[3] A. Agarwal, A. Agarwal, H. Li, and K. Roy, "A single-vt low-leakage gated-ground cache for deep submicron," *IEEE J. Solid-State Circuits*, vol. 38, no. 2, pp. 319–328, 2003.

[4] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr, "Sram design on 65-nm cmos technology with dynamic sleep transistor for leakage reduction," *IEEE J. Solid-State Circuits*, vol. 40, no. 4, pp. 895–901, 2005.

[5] R. Krishnamurthy, A. Alvandpour, G. Balamurugan, N. Shanbhag, K. Soumyanath, and S. Borkar, "A 130-nm 6-ghz 256 × 32 bit leakage-tolerant register file," *IEEE J. Solid-State Circuits*, vol. 37, no. 5, pp. 624–632, 2002.

[6] B. Calhoun and A. Chandrakasan, "A 256kb sub-threshold sram in 65nm cmos," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 6-9, 2006, pp. 2592–2601.

[7] T.-H. Kim, J. Liu, J. Keane, and C. Kim, "A high-density subthreshold sram with data-independent bitline leakage and virtual ground replica scheme," in *Proc. Digest of Technical Papers. IEEE International Solid-State Circuits Conference ISSCC 2007*, 11–15 Feb. 2007, pp. 330–606.

[8] B. Zhai, D. Blaauw, D. Sylvester, and S. Hanson, "A sub-200mv 6t sram in 0.13um cmos," in *Proc. Digest of Technical Papers. IEEE International Solid-State Circuits Conference ISSCC 2007*, 11–15 Feb. 2007, pp. 332–606.

[9] S. P. I. Chang, J. Kim and K. Roy, "A 32kb 10t subthreshold sram array with bit-interleaving and differential read scheme in 90nm cmos," in *ISSCC*, 2008.

[10] N. Verma and A. Chandrakasan, "A 256 kb 65 nm 8t subthreshold sram employing sense-amplifier redundancy," *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 141–149, 2008.

[11] A. Asenov, A. R. Brown, J. H. Davies, S. Kaya, and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale mosfets," *Electron Devices, IEEE Transactions on*, vol. 50, no. 9, pp. 1837–1852, 2003.

[12] A. Bhavnagarwala, X. Tang, and J. Meindl, "The impact of intrinsic device fluctuations on cmos sram cell stability," *IEEE J. Solid-State Circuits*, vol. 36, no. 4, pp. 658–665, April 2001.

[13] R. Heald and P. Wang, "Variability in sub-100nm sram designs," in *Computer Aided Design, 2004. ICCAD-2004. IEEE/ACM International Conference on*, 7-11 Nov. 2004, pp. 347–352.

[14] S. Mukhopadhyay, K. Kim, H. Mahmoodi, A. Datta, D. Park, and K. Roy, "Self-repairing sram for reducing parametric failures in nanoscaled memory," in *VLSI Circuits, 2006. Digest of Technical Papers. 2006 Symposium on*, 2006, pp. 132–133.

[15] F. Hamzaoglu, K. Zhang, and et al., "A 153mb-sram design with dynamic stability enhancement and leakage reduction in 45nm high-k metal-gate cmos technology," in *Proc. Digest of Technical Papers. IEEE International Solid-State Circuits Conference ISSCC 2008*, 3–7 Feb. 2008, pp. 376–621.

[16] Y. Morita, H. Fujiwara, H. Noguchi, K. Kawakami, J. Miyakoshi, S. Mikami, K. Nii, H. Kawaguchi, and M. Yoshimoto, "A vth-variation-tolerant sram with 0.3-v minimum operation voltage for memory-rich soc under dvs environment," in *VLSI Circuits, 2006. Digest of Technical Papers. 2006 Symposium on*, 2006, pp. 13–14.

[17] H. soo Yu, N.-S. Kim, Y.-J. Son, Y.-G. Kim, H.-C. Kim, U.-R. Cho, and H.-G. Byun, "A sram core architecture with adaptive cell bias scheme," in *VLSI Circuits, 2006. Digest of Technical Papers. 2006 Symposium on*, 2006, pp. 128–129.

[18] A. Bhavnagarwala, A. Bhavnagarwala, S. Kosonocky, Y. Chan, K. Stawiasz, U. Srinivasan, S. Kowalczyk, and M. Ziegler, "A sub-600mv, fluctuation tolerant 65nm cmos sram array with dynamic cell biasing," in *Proc. IEEE Symposium on VLSI Circuits*, 2007, pp. 78–79.

[19] M. Khellah, N. S. Kim, Y. Ye, and et al., "Pvt-variations and supply-noise tolerant 45nm dense cache arrays with diffusion-notch-free (dnf) 6t sram cells and dynamic multi-vcc circuits," in *Proc. IEEE Symposium on VLSI Circuits*, 2008, pp. 48–49.

[20] L. Chang, Y. Nakamura, R. Montoye, J. Sawada, A. Martin, K. Kinoshita, F. Gebara, K. Agarwal, D. Acharyya, W. Haensch, K. Hosokawa, and D. Jamsek, "A 5.3ghz 8t-sram with operation down to 0.41v in 65nm cmos," in *Proc. IEEE Symposium on VLSI Circuits*, 14–16 June 2007, pp. 252–253.

[21] K. Flautner, N. S. Kim, S. Martin, D. Blaauw, and T. Mudge, "Drowsy caches: simple techniques for reducing leakage power," in *Computer Architecture, 2002. Proceedings. 29th Annual International Symposium on*, 25-29 May 2002, pp. 148–157.

[22] N. S. Kim, K. Flautner, D. Blaauw, and T. Mudge, "Single-vdd and single-vt super-drowsy techniques for low-leakage high-performance instruction caches," in *Low Power Electronics and Design, 2004. ISLPED '04. Proceedings of the 2004 International Symposium on*, 2004, pp. 54–57.

[23] A. Singhee and R. Rutenbar, "Statistical blockade: A novel method for very fast monte carlo simulation of rare circuit events, and its application," in *Proc. Design, Automation & Test in Europe Conference & Exhibition DATE '07*, 2007, pp. 1–6.

[24] H. Qin, Y. Cao, D. Markovic, A. Vladimirescu, and J. Rabaey, "Sram leakage suppression by minimizing standby supply voltage," in *Quality Electronic Design, 2004. Proceedings. 5th International Symposium on*, 2004, pp. 55–60.

[25] E. Seevinck, F. List, and J. Lohstroh, "Static-noise margin analysis of mos sram cells," *IEEE J. Solid-State Circuits*, vol. 22, no. 5, pp. 748–754, 1987.

[26] B. Calhoun and A. Chandrakasan, "Static noise margin variation for sub-threshold sram in 65-nm cmos," *IEEE J. Solid-State Circuits*, vol. 41, no. 7, pp. 1673–1679, July 2006.

[27] J. Wang, A. Singhee, R. A. Rutenbar, and B. H. Calhoun, "Statistical modeling for the minimum standby supply voltage of a full sram array," in *Proc. ESSCIRC 33rd European Solid State Circuits Conference*, 2007, pp. 400–403.

[28] N. S. Kim, K. Flautner, D. Blaauw, and T. Mudge, "Circuit and microarchitectural techniques for reducing cache leakage power," *IEEE Trans. VLSI Syst.*, vol. 12, no. 2, pp. 167–184, Feb. 2004.

[29] L. Clark, M. Morrow, and W. Brown, "Reverse-body bias and supply collapse for low effective standby power," *IEEE Trans. VLSI Syst.*, vol. 12, no. 9, pp. 947–956, Sept. 2004.

[30] J. Wang and B. Calhoun, "Canary replica feedback for near-drv standby v¡inf¿dd¡/inf¿ scaling in a 90nm sram," in *Proc. IEEE Custom Integrated Circuits Conference CICC '07*, 2007, pp. 29–32.

[31] S. Ghosh, S. Mukhopadhyay, K. Kim, and K. Roy, "Self-calibration technique for reduction of hold failures in low-power nano-scaled sram," in *Design Automation Conference, 2006 43rd ACM/IEEE*, 24-28 July 2006, pp. 971–976.

[32] H. Qin, A. Kumar, K. Ramchandran, J. Rabaey, and P. Ishwar, "Error-tolerant sram design for ultra-low power standby operation," in *Proc. 9th International Symposium on Quality Electronic Design ISQED 2008*, 2008, pp. 30–34.

[33] Y. Ramadass and A. Chandrakasan, "Minimum energy tracking loop with embedded dc-dc converter delivering voltages down to 250mv in 65nm cmos," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2007, pp. 64–587.

[34] Y. Cao, T. Sato, M. Orshansky, D. Sylvester, and C. Hu, "New paradigm of predictive mosfet and interconnect modeling for early circuit simulation," in *Proc. CICC Custom Integrated Circuits Conference the IEEE 2000*, 2000, pp. 201–204.

[35] B. Calhoun and A. Chandrakasan, "Characterizing and modeling minimum energy operation for subthreshold circuits," in *Low Power Electronics and Design, 2004. ISLPED '04. Proceedings of the 2004 International Symposium on*, 9-11 Aug. 2004, pp. 90–95.

[36] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "Theoretical and practical limits of dynamic voltage scaling," in *Proc. 41st Design Automation Conference*, 2004, pp. 868–873.

[37] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Modeling of failure probability and statistical design of sram array for yield enhancement in nanoscaled cmos," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 24, no. 12, pp. 1859–1880, Dec. 2005.

[38] K. Agarwal and S. Nassif, "Statistical analysis of sram cell stability," in *Design Automation Conference, 2006 43rd ACM/IEEE*, 24-28 July 2006, pp. 57–62.

[39] E. Grossar, J. Croon, M. Stucchi, K. Maex, and W. Dehaene, "A yield-aware modeling methodology for nano-scaled sram designs," in *Integrated Circuit Design and Technology, 2005. ICICDT 2005. 2005 International Conference on*, 9-11 May 2005, pp. 33–36.

[40] H. Nho, S.-S. Yoon, S. Wong, and S.-O. Jung, "Numerical estimation of yield in sub-100-nm sram design using monte carlo simulation," *IEEE Trans. Circuits Syst. II*, vol. 55, no. 9, pp. 907–911, 2008.

[41] M. Khellah, D. Khalil, D. Somasekhar, Y. Ismail, T. Karnik, and V. De, "Effect of power supply noise on sram dynamic stability," in *Proc. IEEE Symposium on VLSI Circuits*, 2007, pp. 76–77.

[42] A. Bhavnagarwala, S. Kosonocky, C. Radens, K. Stawiasz, R. Mann, Q. Ye, and K. Chin, "Fluctuation limits & scaling opportunities for cmos sram cells," in *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, 5-7 Dec. 2005, pp. 659–662.

[43] C. Wann, R. Wong, D. Frank, R. Mann, S.-B. Ko, P. Croce, D. Lea, D. Hoyniak, Y.-M. Lee, J. Toomey, M. Weybright, and J. Sudijono, "Sram cell design for stability methodology," in *VLSI Technology, 2005. (VLSI-TSA-Tech). 2005 IEEE VLSI-TSA International Symposium on*, 25-27 April 2005, pp. 21–22.

[44] E. Grossar, M. Stucchi, K. Maex, and W. Dehaene, "Read stability and write-ability analysis of sram cells for nanometer technologies," *IEEE J. Solid-State Circuits*, vol. 41, no. 11, pp. 2577–2588, Nov. 2006.

[45] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng, and M. Bohr, "A 3-ghz 70-mb sram in 65-nm cmos technology with integrated column-based dynamic power supply," *IEEE J. Solid-State Circuits*, vol. 41, no. 1, pp. 146–151, Jan. 2006.

[46] K. Takeda, H. Ikeda, Y. Hagihara, M. Nomura, and H. Kobatake, "Redefinition of write margin for next-generation sram and write-margin monitoring circuit," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2006, pp. 2602–2611.

[47] G. Huang, W. Dong, Y. Ho, and P. Li, "Tracing sram separatrix for dynamic noise margin analysis under device mismatch," in *Proc. IEEE International Behavioral Modeling and Simulation Workshop BMAS 2007*, W. Dong, Ed., 2007, pp. 6–10.